SCIENCE CHINA
Life Sciences

• RESEARCH PAPER •

April 2013 Vol.56 No.4: 324–334
doi: 10.1007/s11427-013-4460-x

Systematic study of human long intergenic non-coding RNAs and their impact on cancer

SUN Liang[1†], LUO HaiTao[2†], LIAO Qi[3], BU DeChao[2], ZHAO GuoGuang[2], LIU ChangNing[2], LIU YuanNing[1*] & ZHAO Yi[2*]

[1]*College of Computer Science and Technology, Jilin University, Changchun 130012, China;*
[2]*Bioinformatics Research Group, Advanced Computing Research Laboratory, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;*
[3]*Department of Preventive Medicine, School of Medicine, Ningbo University, Ningbo 315211, China*

Received October 18, 2012; accepted December 3, 2012; published online March 15, 2013

The functional impact of several long intergenic non-coding RNAs (lincRNAs) has been characterized in previous studies. However, it is difficult to identify lincRNAs on a large-scale and to ascertain their functions or predict their structures in laboratory experiments because of the diversity, lack of knowledge and specificity of expression of lincRNAs. Furthermore, although there are a few well-characterized examples of lincRNAs associated with cancers, these are just the tip of the iceberg owing to the complexity of cancer. Here, by combining RNA-Seq data from several kinds of human cell lines with chromatin-state maps and human expressed sequence tags, we successfully identified more than 3000 human lincRNAs, most of which were new ones. Subsequently, we predicted the functions of 105 lincRNAs based on a coding-non-coding gene co-expression network. Finally, we propose a genetic mediator and key regulator model to unveil the subtle relationships between lincRNAs and lung cancer. Twelve lincRNAs may be principal players in lung tumorigenesis. The present study combines large-scale identification and functional prediction of human lincRNAs, and is a pioneering work in characterizing cancer-associated lincRNAs by bioinformatics.

lincRNA, identification, functional annotation, cancer

Citation: Sun L, Luo H T, Liao Q, et al. Systematic study of human long intergenic non-coding RNAs and their impact on cancer. Sci China Life Sci, 2013, 56: 324–334, doi: 10.1007/s11427-013-4460-x

Long intergenic non-coding RNAs (lincRNAs), a kind of non-coding RNA that were long (>200 nt) and located in the intergenic region of genome, have been paid great attention recently for their regulatory role in many biological processes. Several research teams have already identified a few thousand relatively reliable lincRNAs from mammalian genomes, such as those of human and mouse [1–5]. These lincRNAs had been reported to localize to specific subcellular compartments [6], be involved in numerous regulatory process, exhibit cell type-specific expression [7] and be associated with various diseases. Notably, several lincRNAs have been found to be oncogenic or tumor-suppressor genes and to play important roles in key cancer pathways at transcriptional, post-transcriptional and epigenetic levels [8,9]. However, owing to the complexity of lincRNAs, studying the functions of lincRNAs is difficult. How many lincRNAs are there, what functions do they have, and which are associated with disease or cancer? All these questions remain unanswered.

Current studies of lincRNAs fall into three main categories. The first category includes studies aimed at identifying

†Contributed equally to this work
*Corresponding author (email: biozy@ict.ac.cn; liuyn@jlu.edu.cn)

lincRNAs. Thousands of lincRNAs have been collected into several databases, such as H-inv [10], GENCODE [11], RefSeq [12] and FANTOM [13]. A few studies have identified some relatively reliable lincRNA regions or transcripts based on previous experimental data. For instance, by searching for the chromatin signature of actively transcribed genes [14,15], more than 1600 and 3300 regions containing lincRNAs were identified in the mouse and human genomes, respectively. With the development of sequencing technology, lincRNAs have been identified from RNA-seq data. For example, Guttman et al. [16] identified more than a thousand lincRNAs using software named Scripture, which can reconstruct full-length gene structures from RNA-Seq data. Recently, over 8000 human lincRNAs were assembled from RNA-Seq data across 24 tissues and cell types [17]. Numerous lincRNAs may still be uncovered; however, it is still a challenge to get the complete catalog of the lincRNAs in mammalian genomes.

The second category of studies includes those aimed at performing a functional analysis of lincRNAs. Unlike microRNAs or proteins, the functions of lincRNAs cannot be inferred from their sequences or structure characteristics owing to the diversity of lincRNAs [8]. In addition, functional analysis of lincRNAs is also hampered by the lack of collateral information such as large scale molecular interaction data and expression profiles. Although a few lincRNAs, such as HOTAIR [18], AIR [19], Kcnq1ot1 [20] and lincRNA-p21 [21], have been functionally characterized experimentally, and computational workflows for lincRNA functional annotation have been designed [22], the functions of a large number of lincRNAs remain unknown.

The third category of studies on lincRNAs includes those exploring the links between lincRNAs and diseases or cancer. It has been reported that lincRNAs show differences in expression profiles between normal and tumor samples, and that they have a major role in the development and progression of cancer. For example, in primary breast tumors and metastases, the lincRNA HOTAIR is highly expressed and interacts with polycomb repressive complex 2 (PRC2). This results in an altered pattern of H3K27 methylation of target genes and, thus, an increase in invasiveness and metastasis, whereas the depletion of HOTAIR inhibited cancer invasiveness [23]. Other cancer-associated lincRNAs such as lincRNA-p21, ANRIL and MALAT-1 also have similar roles [24]. However, to date, one study has characterized the association between lincRNAs and cancer [25], and the mechanisms by which lincRNAs affect tumor initiation and/or progression remain poorly understood. To further understand lincRNAs, it is necessary to study human lincRNAs from all three angles systematically.

In the present study, by combining RNA-Seq reads from several kinds of human cell lines with chromatin-state maps and human expressed sequence tags (ESTs), more than 3000 human lincRNAs were successfully identified, almost all of them new ones. Next, we re-annotated the probes of Human Genome U133 Plus 2.0 Array to obtain expression profiles for the newly identified lincRNAs. Then a coding-non-coding gene co-expression network was constructed to determine the functions of the lincRNAs. Finally, we proposed a genetic mediator and key regulator model to explore cancer-associated lincRNAs and their putative mechanisms. As a result, we identified three lincRNAs acting as genetic mediators and twelve acting as key regulators that may play key roles in several important cancer pathways such as the cell cycle and immune processes in lung tumorigenesis. The evidence gathered from this work provides a new approach for computational and systematical analyses of lincRNAs, as well as clues to the functions of lincRNAs and possible roles in cancer.

1 Materials and methods

1.1 Data downloading and processing

Nine human paired-end RNA-Seq data sets spanning six cell types (accession numbers shown in Table S1) were downloaded from the Sequence Read Archive (SRA) database in NCBI (http://www.ncbi.nlm.nih.gov/sra) in September 2010. These RNA-Seq reads, in FASTQ format, consist of one or more runs, each of which has two files for independent left and right reads.

The human reference genome sequence (Hg18) and all RefSeq coding and noncoding genes were downloaded from the University of California Santa Cruz (UCSC) Genome Browser. H-inv transcripts were obtained from UCSC Table Browser. Human ESTs located within protein-coding genes and lincRNA regions as well as their positional data were obtained from the UCSC Table Browser.

We obtained the list of all human lincRNA regions in the human genome (Hg18) along with mouse lincRNA regions as determined by Guttman et al. [16] in the mouse genome (MM8), and used the liftOver (http://genome.ucsc.edu/cgi-bin/hgLiftOver) tool to identify their orthologous locations in the human genome (Hg18).

Microarray data sets were downloaded from the Gene Expression Omnibus (GEO) database. Accession numbers of data sets used are GSE7216, GSE12161, GSE6241, GSE8884, GSE10313, and GSE10021 (Table S2).

1.2 Identification of lincRNAs

First, experiment reads were aligned to the human reference genome (Hg18) using the TopHat aligner. Then, Scripture was run separately according to the walkthrough described on the website (http://www.broadinstitute.org/software/scripture/Walkthrough_example). For each experiment, the final Scripture results included BED files containing all reconstructed transcripts and the expression levels of these transcripts, as well as DOT files containing a transcript graph constructed within Scripture. Second, we compiled the re-

sults of the nine experiments into a single file containing information about all of the transcripts reconstructed, using the findpeaks method (setting parameters "-dist_type 3 -subpeaks 0.2 -minimum 1 -trim 0.65"). Third, we subtracted all transcripts mapping to exons, introns and the antisense transcripts overlapping the protein-coding genes and eliminated those less than 200 bp in length and within 1 kb of protein-coding genes.

To construct EST-origin transcripts, we extracted all ESTs located within either protein-coding genes or lincRNA regions. Next, within each gene unit, we used the findpeaks method to identify EST peaks using the parameters "-dist_type 3 -subpeaks 0.2 -minimum 5 -trim 0.25". After trimming and filtering, we join the peaks by turns in the same gene location according to the orientation of ESTs.

We combined Scripture-origin and EST-origin noncoding transcripts. To measure coding potential capacity, all noncoding transcripts and protein-coding gene sequences were submitted in FASTA format to CNCI software (in house) and the CPC web server (http://cpc.cbi.pku.edu.cn/). Based on the CNCI and CPC score distribution of coding and noncoding genes, we removed noncoding genes with a CNCI score above a threshold of 10 and those with a CPC score above 0. To estimate the evolutionary constraints among mammalian sequences we constructed the cumulative distribution of PhastCons scores for introns, RefSeq coding genes and our lincRNAs. To detect the gene transcriptional signature, we detected the H3K4me3 and H3K36me3 peaks generated by the Broad/MGH ENCODE group in gene promoter and gene body regions, respectively.

1.3 MSD value

Mean similarity degree (MSD) values were used to evaluate the strength of the sequence similarity between two transcripts, defined as $MSD=(L_{overlap}/L_{T1}+L_{overlap}/L_{T2})/2$, where L_{T1} and L_{T2} are the length of two transcripts, $L_{overlap}$ is the length of the overlapping regions between them.

1.4 Findpeaks method

Findpeaks is a piece of java software developed by Fejes, which can identify areas of fragment enrichment. It is available at "http://www.bcgsc.ca/platform/bioinfo/software/findpeaks". First, the bed file of all fragments was separated by chromosome, and sorted by read position. Next, we set up the findpeaks parameters such as subpeaks and trim at the appropriate thresholds, and ran Findpeaks.

1.5 Probe re-annotation

We used probe re-annotation pipeline as previously described [22]. Briefly, probe sequences for the Affymetrix Human Genome U133 Plus 2.0 Array downloaded from the Affymetrix website were aligned to our lincRNA sequences and to the RefSeq coding transcript sequences, respectively, using BLASTn. Only perfectly matched probes and transcripts were maintained, resulting in two sets of probes targeting protein-coding and non-coding transcripts. We removed probes in the non-coding probe set that perfectly matched RefSeq coding sequences. Next, we used the Entrez GeneID as an identifier of coding genes, and mapped the protein-coding probes from the transcription level to the gene level. Finally, we removed the probes that matched more than two genes or non-coding transcripts and selected the genes and transcripts that have at least three probes. A new CDF package (called re-annotated hgu133plus2) covering the re-annotated probe-gene relationships was created using R package.

1.6 Comparison of gene expression as measured by RNA-Seq and re-annotated Human Genome U133 Plus 2.0 Array

RNA-Seq data were used to measure the expression levels of transcripts in conjunction with Scripture software, which computed reads per kilobase of exon model per million mapped reads (RPKM). Preprocessing of microarray data was done using R Bioconductor software and consisted of Robust Multichip Average (RMA) background correction, constant normalization and expression summarization as described by Liao et al. [22]. Expression intensity was $\log_2$-transformed. For each cell line, Spearman correlation coefficients for the same genes in the two data sets were calculated. For each non-coding gene, the correlations among the expression levels across six cell lines in the two data sets were also calculated. As a control, non-coding genes were randomly paired and Spearman correlation coefficients were computed. The control step was repeated 1000 times.

1.7 Gene correlation and construction of the co-expression network

Thirty-one data sets were used to construct the coding-non-coding gene co-expression network. For each data set, the data processing was similar to the workflow previously described [22]. Briefly, genes with expressional variance ranked in the top 75th percentile of each data set were retained. Second, a set of Pearson correlation coefficient (Pcc) *P*-values for each gene pair was estimated using Fisher's asymptotic test and adjusted using the Bonferroni multiple test correction. Only gene pairs with a *P*-value $\leqslant 0.01$ and with a Pcc value ranked in the top or bottom 0.5 percentile for each gene were regarded as co-expressed in the given data set. Third, each gene pair was assigned a score according to the number of data sets in which the gene pair was co-expressed in the same 'direction' (i.e., positively or negatively). We constructed several networks using different thresholds (the number of data sets) and measured the topological properties of the resulting networks. All of the above

processes, including determining the clustering coefficient, gamma, and scale-free topology criteria, were implemented in R software.

1.8 Random networks

Random networks were constructed as previously described [22]. Briefly, random networks consisted of the same genes as in the observed networks and have the same topological properties as observed networks. To do this, a gene in the random network had the same number of connections as in the observed network, but its links to other genes were random instead of being based on co-expression patterns.

1.9 Enrichment analysis

Enrichment analysis of Gene Ontology Biological Processes as well as KEGG pathways were performed using the g:profiler web server with the default parameters.

1.10 Genetic mediators of cancer

The reverse-engineered networks used the mode-of-action by network identification (MNI) algorithm, which involved two phases as described previously. In phase one, we used a total of 1037 microarray expression profiles spanning several different cancer types as a training set to reverse engineer a coding-noncoding gene regulatory network. In phase two, we used the expression profiles of lung cancers and normal samples as a test set and the network as a filter to determine the genes affected by these conditions. The MNI algorithm software was downloaded from the website "http://gardnerlab.bu.edu". The parameters used with the MNI algorithm software are as follows: the threshold for significant external influence (thP) was set to 0.25; the fraction of genes to be kept (Kfrac) was set to 0.33; and the number of rounds (NROUNDS) was set to 3.

1.11 The cancer gene set map

We obtained gene expression profiles for five data sets including 546 arrays spanning four cancer types and normalized the expression of each gene in every data set separately. First, the expression value of each gene g was $\log_2$-transformed (truncating to an expression value of 10 any below that value). Second, we normalized the ($\log_2$-transformed) expression value of gene g in each array relative to its average expression in all the arrays in the same data set, by subtracting its average expression value in that data set. We then extracted coding and non-coding gene centered gene set from our co-expression network, which consists of a central gene that have our functional annotations and coding genes that connect directly to it. By creating a gene set map, we used Genomica software, which can characterize an expression dataset on the basis of gene sets and experiment sets that significantly change within it. Genomica download and workflow are available at http://genomica.weizmann.ac.il/Tutorial/create_module_map.html.

2 Results

2.1 Integrated method to identify human lincRNAs

In the method based on the RNA-Seq data, nine human RNA-Seq experimental data sets spanning six cell types downloaded from ENCODE project [26] were used to construct transcripts using Scripture [16] (Table S1). Then, nine sets of transcripts were combined and 19640 non-overlapping, multi-exonic transcripts called as scripture-origin transcripts were obtained, among which 16972 corresponded to RefSeq protein-coding genes (Figure S1). The MSD value, which was defined to evaluate the strength of sequence similarity of two transcripts (Materials and methods), was high between these transcripts and corresponding RefSeq coding genes (Figure 1A), indicating that the transcripts constructed by Scripture were of good quality. Then, all scripture-origin transcripts that were mapped to exons, introns and the antisense transcripts of protein-coding genes were subtracted. In addition, transcripts less than 200 nt in length and those within 1 kb of the protein-coding genes were eliminated to avoid promoter and 3′-associated transcripts. The pipeline led to 1746 putative lincRNAs with an average length of 1.4 kb, and with 3.9 exons of 355 bp on average, among which 449 overlapped with the 3000 distinct lincRNA loci that were previously identified based on the K4-K36 regions described by Khalil et al. [15].

Then, we constructed new lincRNA gene structures using a method based on expressed sequence tags (ESTs), which were usually used to discover new transcript models (Figure S1). All human ESTs located in both protein-coding gene and lincRNA regions were used to find EST-peaks in conjunction with FindPeaks software [27]. Using an appropriate threshold, the left EST-peaks were linked to the corresponding genes. Finally, 18959 transcripts corresponding to 18712 coding genes and 1995 lincRNA regions were constructed and named as EST-origin transcripts. The quality of these transcripts was also high, with high MSD values for EST-origin transcripts and their corresponding RefSeq protein-coding transcripts (Figure 1A). On average, the 1995 lincRNA regions were 1.6 kb long, and had 4.5 exons of 356 bp. In addition, 317 overlapped with 341 scripture-origin lincRNAs in the genome region.

Combining the 1746 scripture-origin lincRNAs with 1995 EST-origin lincRNAs, 3528 putative lincRNAs in human genome were identified in total, among which 563 had been annotated as RefSeq noncoding RNAs. Evaluating the coding potential of each unannotated lincRNA using Coding-Non-Coding Index (CNCI) software (in house) and a Coding Potential Calculator [28], we found that almost

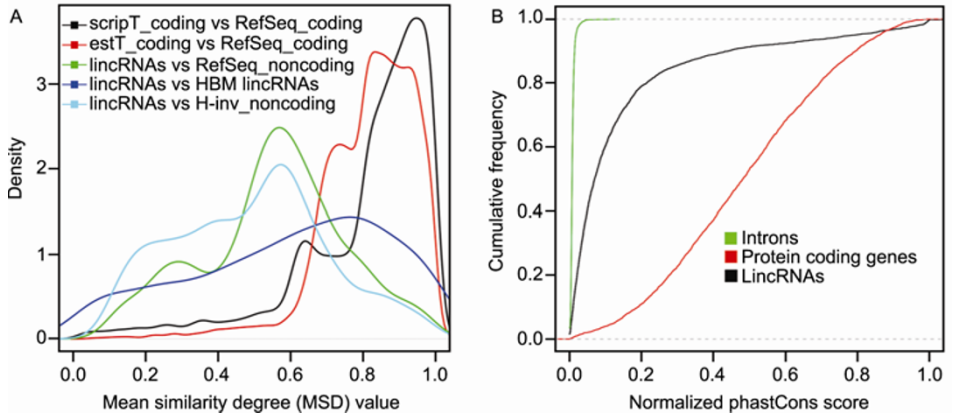

Figure 1 Integrated method to identify human lincRNAs. A, Distribution of mean similarity degree (MSD) values between scripture-origin coding genes and RefSeq coding genes (black), EST-origin coding genes and RefSeq coding genes (red), our identified lincRNA genes and RefSeq non-coding genes (green), our identified lincRNA genes and Human Body Map (HBM) lincRNA genes (blue), and our identified lincRNA genes and H-inv non-coding genes (cyan). B, Conservation of the genomic transcript sequences for lincRNAs, protein-coding genes and introns.

90% of them were predicted as non-protein-coding genes. After eliminating those transcripts that were predicted as coding genes, 3215 putative lincRNAs were left; on average, these were 1.5 kb long, with 4.1 exons of 358 bp (File S1). Among the 3215 lincRNAs, 17% overlapped with RefSeq noncoding genes, 27% overlapped with H-inv transcripts [10], while 12% overlapped with Human Body Map lincRNAs [17]. MSD values for comparisons between our identified lincRNAs and the above three types were high (Figure 1A), indicating the good quality of these lincRNAs. Furthermore, we compared our lincRNAs catalog with transcripts of GENCODE V12 identified by the ENCODE project [29], which has just published 30 papers including a few that extensively characterize lincRNAs, and found that 47% (1502/3215) of our lincRNAs overlapped with GENCODE transcripts in terms of genomic location. The sequence conservation of the lincRNAs was lower than that of protein-coding genes but higher than that of intron regions (Figure 1B). Moreover, chromatin signature analysis revealed that >60% of these lincRNAs had both H3K4me3 peaks in the promoter region and H3K36me3 peaks in the gene body, implying that lincRNAs have a similar transcriptional signature to protein-coding genes [1,2]. Altogether, using RNA-Seq and EST data sets we identified, in total, 3215 human lincRNAs, among which nearly 80% were novel.

2.2 Expression profiles of lincRNAs

To analyze the expression of lincRNAs, we used our previously well-established computational workflow to re-annotate the probes of Human Genome U133 Plus 2.0 Array [22]. As a result, 297510 and 5731 of 604258 probes were mapped to 16883 protein-coding RNAs and 492 lincRNAs, respectively, and were used to assemble a new chip-description file. To evaluate the accuracy of the re-annotated array, we compared the expression levels of the mRNAs and lincRNAs detected based on RNA-Seq data with those detected by the re-annotated Human Genome U133 Plus 2.0 Array. The results showed that the expression levels of both mRNAs and lincRNAs from the two independent datasets had a significantly high level of correlation (*P*-value of the KS test were less than 10^{-10}) (Figure S2A). Furthermore, the two independent datasets detected several of the same cell type-specific lincRNAs (Figure S2B).

2.3 Predicting lincRNA gene functions using co-expression networks

Next, a coding-non-coding gene co-expression network was constructed and used to determine lincRNA functions by the methods described previously [22]. Thirty-one microarray datasets involving a number of biochemical and biophysical conditions, various tissue resources, and diverse biological processes from GEO were used to construct the co-expression network (Table S2). The last co-expression network was composed of the gene pairs that were co-expressed in at least a certain cutoff of microarray datasets. To obtain a co-expression network with high quality, we evaluated the topological properties of a series of networks obtained by different cutoffs (edges between two genes were included in the network only if the two genes were co-expressed in the same direction in more than a given number of datasets) (Table S3). GO term overlap analysis showed that the higher the cutoff, the more similar were the annotated functions of neighboring genes in the network (Table S3). Based on the topologies of the networks, we focused on the co-expression network that was constructed with a cutoff of 5 to obtain accurate topological and biological properties. The resulting network included 256 lincRNA genes and 10802 coding genes, with 48946 coding-coding edges, 2721 coding-noncoding edges and 79 noncoding-noncoding edges (Table S3).

The ‘two-color’ co-expression network had a scale-free

topology, which satisfies the required characteristics of biological networks [30] (Table S3). For example, our network comprised many genes with relatively few connections (mean=9, median=4), but a few genes, also called as hubs, had large number of connections and could influence the expression of many other genes. The gene pairs in the co-expression network shared many Gene Ontology (GO) annotations ($P<10^{-16}$), more than did those in the random network (Table S3). Of the 48946 coding-coding gene pairs, 32740 pairs had identical GO annotations, among which 6536 (20%) had the same GO annotations, while only 5% in the random network did. These results suggested that our co-expression network had appropriate topological and biological properties.

To obtain the functional characteristics of lincRNAs, two different methods including gene-centered and co-expressed module sub-networks [22,31] were used to predict functions. First, we parsed the co-expression network into 4688 subnetworks, each of which consisted of a central gene and its directly connected genes. Among them, 114 were lincRNA-centered while 4574 were coding gene-centered subnetworks. Then, we calculated the enriched GO terms for each subnetwork and found that 3054 (65%) (2954 coding gene-centered and 100 lincRNA-centered subnetworks) had at least one significantly enriched GO term. Of the 2954 central coding genes, 1112 (38%) had the same GO terms as previously annotated; by contrast, only 188/2761 (7%) in the random ($P<10^{-16}$) (2716 coding gene-centered subnetworks were found in random network). These results demonstrated that the predicted functions of 100 central lincRNAs were relatively reliable. Second, we used the MCL algorithm to parse the network into 2872 modules with five or more genes, of which 209 modules had at least one enriched GO term; 21 of these were composed of both coding and noncoding genes, involving 23 lincRNAs, among which 18 had also been predicted by the first method and 13 had the same predicted functions. Altogether, 105 lincRNAs' functions were predicted using the two above methods (File S2). For example, TUG1 (named Both_114 in our catalog of lincRNA genes) may serve as a downstream transcriptional repressor in the p53 pathway to repress cell-cycle progression in response to DNA damage [15]. To further validate the predicted functions of lincRNAs, we analyzed the expression characters of lincRNAs through the re-annotated expression profiles of three cell lines treated with a cyclin-dependent kinase (CDK) inhibitor. In total, 1254 coding genes and 39 lincRNA genes were differentially expressed in a dose-responsive manner to the CDK inhibitor, and these were shared by all three cell lines (Figure S3A and B). GO biological process enrichment analysis (Figure S3C) revealed that the functions of 1254 coding genes were significantly enriched for "cell cycle" and "response to DNA damage stimulus" pathways, consistent with evidence that CDK inhibitors cause dysregulation of the cell cycle [32]. Among the 39 differentially expressed lincRNAs, 16 had functional annotation and most of them have regulatory functions. Another lincRNA named Est_835, whose function was predicted as cell cycle regulation, was differently expressed following CDK inhibitor treatment, suggesting that it may be involved in the cell cycle pathway (File S2).

2.4 Predicting the roles of lincRNA genes implicated in cancer

To study the relationship between lincRNAs and cancer, we proposed a genetic mediator and key regulator model, which can identify genes directly affected by cancer or which play key roles in tumorigenesis. The description of the model was as follows. We considered that cancer was caused by the dysregulation of the genes involved in a cancer-related network (CRN), which consisted of a variety of connections such as protein-protein interactions and regulatory interactions, and which may be responsible for tumor occurrence, growth and metastasis. Although hundreds to thousands of genes show different expression levels in cancer compared with normal samples, only two types of genes in the CRN are important for tumorigenesis. One is genes whose different expression between tumor and normal samples is the main reason for dysregulation of cellular pathways. These genes are usually directly affected by cancer and located in the top layer of the CRN, and are called genetic mediators of cancer (Figure 2A). Another type is hubs of the CRN, whose alteration leads to the dysregulation of their numerous regulated genes and which act as key regulators during cancer development (Figure 2A).

We applied the genetic mediator and key regulator model to lung cancer and found several genetic mediators. First, we identified genes, including both coding and lincRNAs, that were differentially expressed in cancer, and further obtained genetic mediators via reverse-engineered gene regulatory networks, which had been verified as effective in prostate cancer [33,34]. The reverse-engineered network is a directed graph that reflects the contribution of each gene to the others, and in which the edge represents how the activity of one gene (genetic mediator) influences the transcription of another gene. Using this method, we obtained 100 top genetic mediators including three lincRNAs specific to lung cancer (Table S4, Figure S4). GO enrichment analysis showed that the functions of these cancer-specific genetic mediators were mainly cell proliferation and signal transduction (Figure 2B). This result was consistent with the previous conclusion that most of the mutated genes in lung adenocarcinoma were involved in several important signaling pathways, such as cell proliferation, cell death and the cell cycle [35]. To further validate our predictions, we collected 2897 lung cancer-related genes, which were extracted manually from the literature or which showed occurrence of somatic mutations in lung cancer, and found that 42 of 97 coding genetic mediators had been experimentally validated

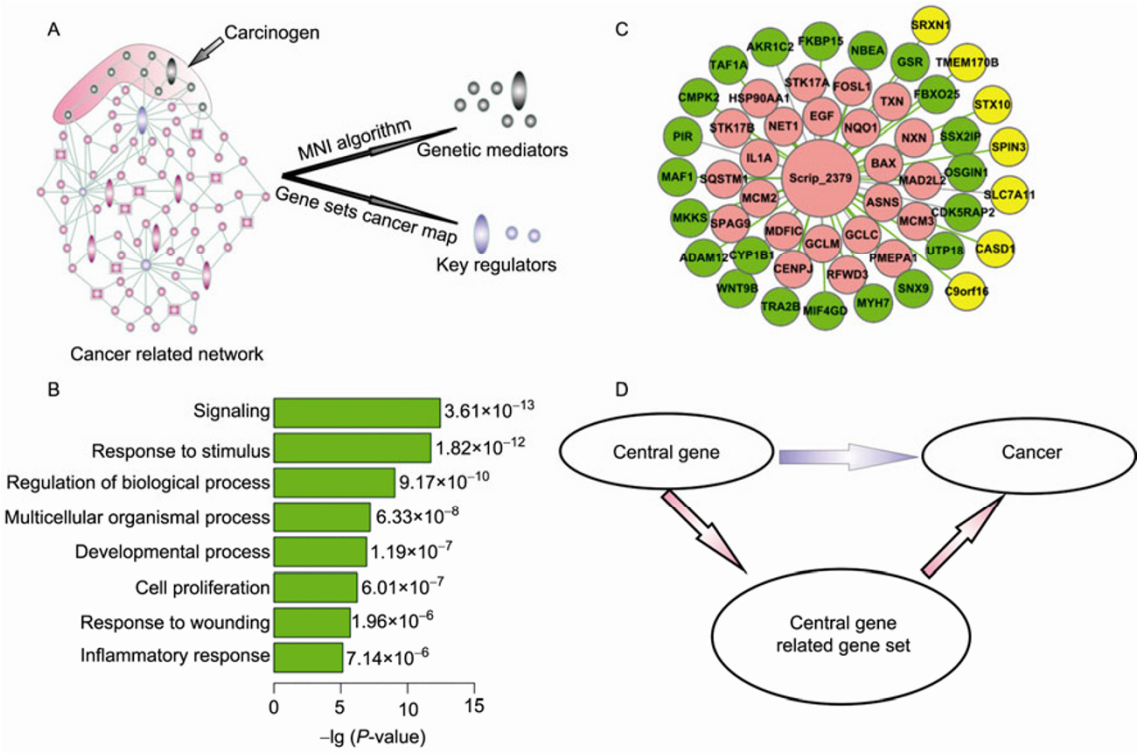

Figure 2 LincRNA genes implicated in cancer. A, Schematic representation of the characterized cancer-associated lincRNAs. Genes disturbed directly by carcinogens are genetic mediators of cancer and at the top layer of the cancer network (black nodes). Genes acting as hubs in the cancer network are key regulators (purple nodes). Rounded nodes denote coding genes; elliptic nodes denote long ncRNA genes. B, GO enrichment analysis revealed 97 coding genes as genetic mediators in cancer. C, LincRNA Scrip_2379 subnetwork. This subnetwork consists of Scrip_2379 (center) and its 51 direct neighbors. Genes colored in pink are known to play a role in the cell cycle and signaling pathways. Genes colored in green are involved in the regulation of cellular process. Yellow denotes other functions. Edges colored in light grey (or green) are negative (or positive) correlations. D, Schematic representation of the characterized key regulators in cancer. It is hypothesized that if a central gene has a close relationship with its gene set, which is associated with cancer, this central gene could be associated with cancer via the function of the gene set.

as lung cancer-related genes ($P<10^{-16}$). Among the three lincRNA genetic mediators, the functions of Scrip_2379 (Figure 2C) and Scrip_2616' were annotated as cell cycle regulation. Overall, using the reverse-engineered coding-noncoding gene regulatory network, we identified three lincRNAs located in the top layer of the regulatory network and which acted as genetic mediators with important roles in cancer.

Next, we characterized lincRNAs acting as key regulators in tumorigenesis and their putative mechanisms implicated in cancer. We presumed that if a gene was closely related to a gene set associated with cancer, the gene may be associated with cancer (Figure 2D). We found that the functional similarity of the central gene and its directly co-expressed genes was high (the ratio might reach 50%) when the number of its co-expressed genes was >20. Therefore, we considered that such a central gene and its directly co-expressed gene set had a close relationship. Then, we investigated the linkage between gene sets and cancer by drawing a cancer map based on the gene sets (we named this the cancer gene set map, CGSM), which could determine whether the sets were induced or repressed in a significant fraction of the microarrays with cancer condition [36]. In the co-expressed network, 73 lincRNAs-related gene sets were obtained and 500 coding gene-related gene sets were randomly selected. Then, we measured the relationships between these gene sets and cancer through 546 microarray expression profiles involving four cancer types and corresponding normal samples (Figure 3A). Several gene sets had similar functions and were combined into the same cluster. By CGSM analysis of 500 coding gene-related gene sets, we found that clusters related to immune and the inflammatory response were repressed in lung tumors, while those related to the cell cycle were induced across several cancer types (Figure 3A). The results were consistent with the hypothesis that tumors were caused by aberrations in the regulation of key immune system and survival pathways. In the results, 167 coding gene-centered gene sets were differently expressed in lung cancer and many of these were genes annotated with the cell cycle and immune system processes (Figure S5A). Thus, these central genes were regarded as putative lung cancer-related genes. Comparing these with the 2897 lung cancer-related genes we collected, we found that a significant portion (79/167) were the same ($P<10^{-16}$). For example, the KIF11-related gene set, in which 56 of 80 genes were involved in the cell cycle pro-

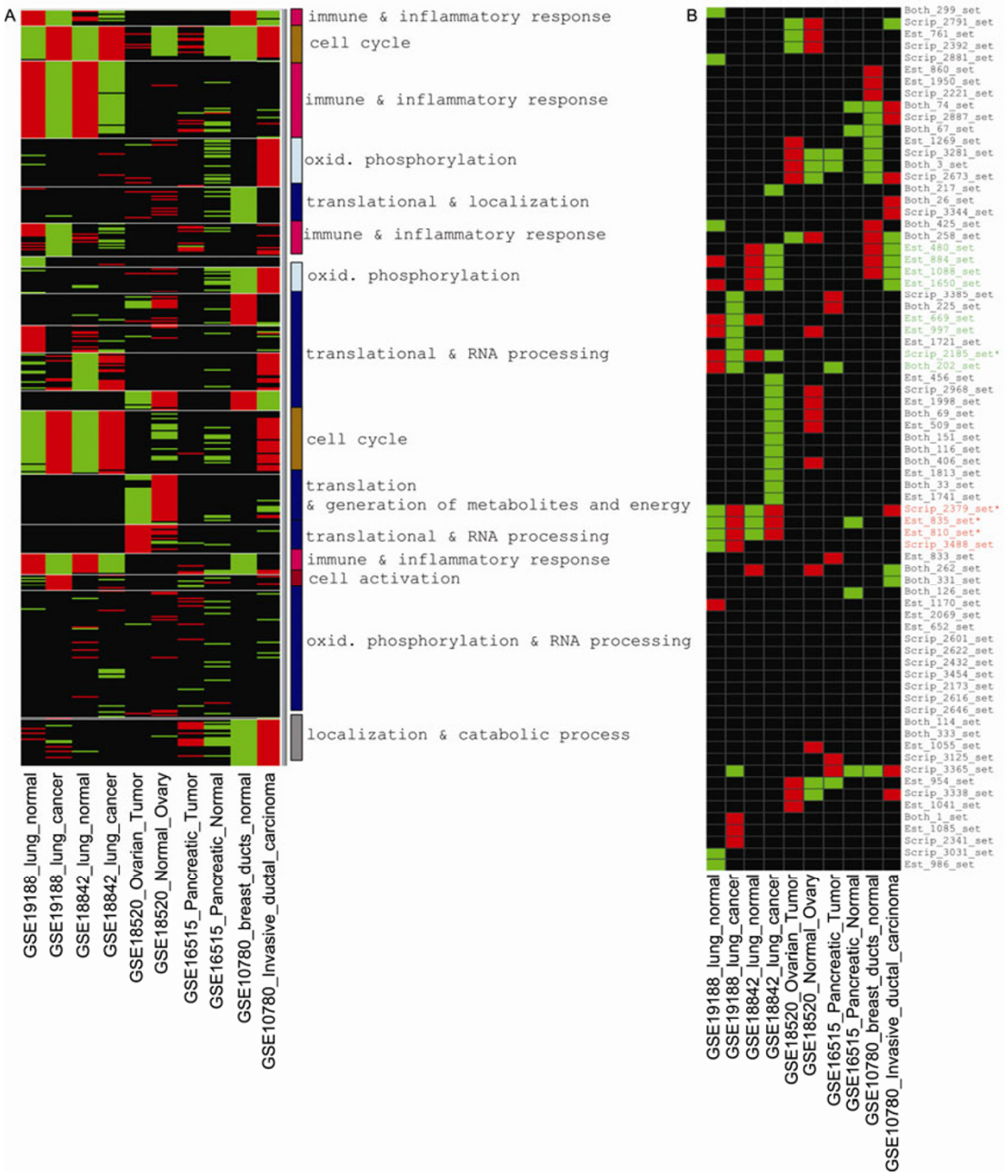

Figure 3 The cancer gene set map (CGSM). A, The CGSM of coding gene-centered gene sets. The rows indicate the state of each gene set and the columns indicate experiment sets. The red (or green) entry indicates that the experiment sets in which the corresponding gene set was significantly induced (or repressed) contained more arrays enriched for the given gene set than would be expected by chance. The rows were clustered into distinct clusters, and the resulting clusters are indicated by horizontal lines. We manually assigned a concise label to each cluster (right; colored bars). B, The CGSM of lincRNA gene-centered gene sets. The stars denote sets detected in both experiments.

cess, could be induced in several cancer types (Figure S5B). Previous reports have shown that KIF11, acting as an attractive anticancer target and playing a key role during cell mitosis, is associated with several types of tumors such as non-small cell lung cancer, glioblastomas, and breast cancer [37]. These results suggested that we can identify new cancer-related genes through exploring the relationships between gene-related gene sets and cancer.

We then used this approach to explore cancer-related lincRNAs. First, the expression of 73 lincRNA-centered gene sets (lincRNA-sets) across various conditions was detected; among these, some were induced or repressed in cancer (Figure 3B). For example, there were 12 lincRNA-sets dysregulated in lung cancer relative to normal lung, and four of them were detected in both experiments (Figure 3B, Table S5). We proposed that the 12 central lincRNAs may be associated with lung cancer. As an example, the Scrip_2379-related gene set, including 51 coding genes the

functions of which were annotated as negative regulators of cell death or apoptosis and positive regulators of the cell cycle, was significantly up-regulated in lung cancer ($P< 10^{-8}$). It is noteworthy that Scrip_2379 also acted as a genetic mediator in lung cancer (Figure 2C). We thus presumed that this lincRNA may have an impact on lung cancer through the cell cycle process. As another example, the Scrip_2185-related set, including 20 coding genes the functions of which were annotated as immune and T cell activation-related functions, was significantly down-regulated in lung cancer ($P<10^{-10}$) (Figure S6A). Interestingly, the genomic location of lincRNA Scrip_2185 overlapped with several T cell receptor beta chain-related noncoding genes (Figure S6B). These results suggested that the metabolic disturbance and signal pathway dysregulation, both of which had been validated as the cause of cancer, may also be associated with lincRNAs (Table S5). Therefore, lincRNAs may be used for cancer diagnosis and prognosis, and as potential therapeutic targets.

3 Discussion

LincRNAs were first discovered by large-scale sequencing of full-length cDNA libraries in mouse and expression profiling using high-resolution genome tiling arrays [38]. RNA-Seq has allowed the analysis of mammalian transcriptomes with an unprecedented resolution and opened the way to studying lincRNAs in mammalian cells [39]. Numerous methods have been established based on RNA-Seq [40–43]. However, most of them depend on existing gene annotations, and thus they have been limited to the discovery of new lincRNAs and their complete structures. Therefore, we applied an alternative method, Scripture, which uses an *ab initio* reconstruction approach to identify the complete transcriptome of an individual sample solely from the unannotated genome sequence and RNA-Seq reads [16]. Multiple studies have shown that significant numbers of lincRNAs exhibit cell type-specific expression [9] and localize to specific subcellular compartments. The catalog of human lincRNAs genes is certain to be incomplete, because it is based on RNA-Seq reads of only six cell types. Using a completely different method, Khalil et al. [15] recently identified ~3300 human lincRNAs by analyzing the chromatin-state maps of six human cell types. However, the structures of most of these genes are not available. To reliably expand our catalog of human lincRNAs as much as possible, we extracted ESTs located in lincRNA regions and used the findpeaks method to construct thousands of putative lincRNA transcripts. Then, we only focused on intergenic lincRNAs while other kinds of ncRNAs, such as antisense transcripts, promoter-associated transcripts, 3′ UTR-associated transcripts and intronic transcripts of protein coding genes, were removed to avoid the extension of or transcriptional noise from protein-coding transcripts [44]. The median distance of lincRNA to neighboring protein-coding gene was >100 kb further indicating that their transcriptions were independent. Finally, after applying a series of filters, we obtained >3000 lincRNAs, most of which had transcriptional signatures and higher evolutionary conservation relative to introns. In fact, although two different methods and large data sets were used to identify lincRNAs, the real repertoire of these transcripts in human cells contains many more transcripts than those cataloged in this study.

For large-scale functional annotation of these lincRNAs, we applied a previously established workflow [22]. The workflow included two major steps: re-annotation of the Human Genome U133 Plus 2.0 Array and construction of a "two color" co-expression network. Unlike previous work, which predicted the functions of lincRNAs using three methods, we did not apply a method based on genomic co-location and only used a hub-based and module-based method in this study, because the median distance of lincRNAs to their corresponding neighboring protein-coding gene was large (>100 kb). The results obtained from the two functional prediction methods were coherent and complementary, strengthening the validity of the predictions.

It is difficult to characterizing the association between lincRNA and cancer by experiments owing to the complexity and scarcity of information on lincRNAs. A few studies have shown that some lincRNAs, such as HOTAIR, lincRNA-p21 or MALAT-1, acting as "tumor-suppressor ncRNAs" or "oncogenic ncRNAs", have a major role in the development of cancer [21,23,45]. For example, lincRNA-p21 could be located within the promoters of p53-regulated genes through a physical association with hnRNP-K to mediate gene expression in a p53-mediated apoptosis pathway, and was considered as a repressor in p53-dependent transcriptional responses [21]. This study implied that lincRNAs may serve as tumor suppressor genes or oncogenes to play important roles in cancer, and that they may be used as potential targets for the development of cancer therapies in the future.

Cancer is an extremely complex disease mainly caused by the mutational alteration of numerous genes including both coding and noncoding genes that control critical cellular pathways. Thus, a deep understanding of cancer will require a comprehensive characterization of the genes that are dysregulated in tumors, which cannot be achieved solely by experiment. The most common method for large-scale identification of tumor-associated genes is comparing the expression profiles of tumor with those of normal samples. However, there are usually hundreds to thousands of genes that exhibit expression changes in cancer relative to normal samples. Thus, it is important to determine the genes acting as principal players in tumorigenesis. To address this problem, we proposed a genetic mediator and key regulator model, which can determine the genes including both protein coding genes and lincRNAs that are directly affected by

cancer or which play key roles in tumorigenesis. To identify genetic mediators of cancer, Collins et al. developed a method and applied it to non-recurrent primary and metastatic prostate cancer data [33,34]. In the present study, we identified three lincRNAs as genetic mediators in lung cancer using this method. Our results also show that dysregulation of several important signaling pathways such as cell proliferation, cell death and the cell cycle is involved in lung cancer. It is noteworthy that except for genetic mediators at the top layer of gene regulatory network, a few genes such as hubs relating with many other genes also play key roles during cancer development. On this point, we used lincRNA-centered subnetworks parsed from our co-expression network as central lincRNA-related gene sets. We proposed that if gene expression of the gene sets changed significantly in cancer, the central gene could associate with cancer. Based on the expression of central genes and their gene sets, we drew a cancer map showing whether the sets were induced or repressed in cancer [36]. According to this analysis, we identified 12 lincRNA-gene sets that were dysregulated in lung cancer; therefore, the central genes may be core regulators in tumorigenesis.

Our study is the first attempt using combinational methods to systematically analyze human lincRNAs, offering a new approach to characterizing cancer-associated lincRNAs by bioinformatics. The results of our work will pave the way for computational analysis of lincRNAs and will be a valuable resource for further biological research. Just as miRNAs act as critical gene regulators in cells and as effective targets in cancer therapeutics, the majority of lincRNAs we identified also have various functions in cellular networks and may provide new approaches to the diagnosis and treatment of cancer. Nevertheless, the findings described in this article are merely a starting point for the study of lincRNAs, and much effort should be put into unveiling the mystery of lincRNAs.

This work was supported by Beijing Natural Science Foundation (5122029).

1 Guttman M, Amit I, Garber M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature, 2009, 458: 223–227
2 Khalil A M, Guttman M, Huarte M, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci USA, 2009, 106: 11667–11672
3 Cabili M N, Trapnell C, Goff L, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev, 2011, 25: 1915–1927
4 Trapnell C, Williams B A, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol, 2010, 28: 511–515
5 Guttman M, Garber M, Levin J Z, et al. *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol, 2010, 28: 503–510
6 Sone M, Hayashi T, Tarui H, et al. The mRNA-like noncoding RNA Gomafu constitutes a novel nuclear domain in a subset of neurons. J Cell Sci, 2007, 120: 2498–2506
7 Mercer T R, Dinger M E, Sunkin S M, et al. Specific expression of long noncoding RNAs in the mouse brain. Proc Natl Acad Sci USA, 2008, 105: 716–721
8 Mercer T R, Dinger M E, Mattick J S. Long non-coding RNAs: insights into functions. Nat Rev Genet, 2009, 10: 155–159
9 Wilusz J E, Sunwoo H, Spector D L. Long noncoding RNAs: functional surprises from the RNA world. Genes Dev, 2009, 23: 1494–1504
10 Yamasaki C, Murakami K, Fujii Y, et al. The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. Nucleic Acids Res, 2008, 36: D793–799
11 Harrow J, Denoeud F, Frankish A, et al. GENCODE: producing a reference annotation for ENCODE. Genome Biol, 2006, 7: S41–49
12 Pruitt K D, Tatusova T, Maglott D R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res, 2007, 35: D61–65
13 Bono H, Kasukawa T, Furuno M, et al. FANTOM DB: database of Functional Annotation of RIKEN Mouse cDNA Clones. Nucleic Acids Res, 2002, 30: 116–118
14 Guttman M, Amit I, Garber M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature, 2009, 458: 223–227
15 Khalil A, Guttman M, Huarte M, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci USA, 2009, 106: 11667
16 Guttman M, Garber M, Levin J Z, et al. *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol, 2010, 28: 503–510
17 Cabili M N, Trapnell C, Goff L, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev, 2011, 25: 1915–1927
18 Rinn J, Kertesz M, Wang J, et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell, 2007, 129: 1311–1323
19 Nagano T, Mitchell J A, Sanz L A, et al. The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. Science, 2008, 322: 1717–1720
20 Pandey R R, Mondal T, Mohammad F, et al. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol Cell, 2008, 32: 232–246
21 Huarte M, Guttman M, Feldser D, et al. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell, 2010, 142: 409–419
22 Liao Q, Liu C, Yuan X, et al. Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. Nucleic Acids Res, 2011, 39: 3864–3878
23 Gupta R A, Shah N, Wang K C, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature, 2010, 464: 1071–1076
24 Huarte M, Rinn J L. Large non-coding RNAs: missing links in cancer? Hum Mol Genet, 2010, 19: R152–161
25 Chen G, Wang Z, Wang D, et al. LncRNADisease: a database for long noncoding RNA associated diseases. Nucleic Acids Res, 2013, 41: D983–986
26 Bernstein B E, Birney E, Dunham I, et al. An integrated encyclopedia of DNA elements in the human genome. Nature, 2012, 489: 57–74
27 Fejes A P, Robertson G, Bilenky M, et al. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics, 2008, 24: 1729–1730
28 Kong L, Zhang Y, Ye Z Q, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res, 2007, 35: W345–349
29 Dunham I, Kundaje A, Aldred S F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature, 2012, 489: 57–74
30 Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol, 2005, 4: Article17
31 Nayak R R, Kearns M, Spielman R S, et al. Coexpression network

based on natural variation in human gene expression reveals gene interactions and functions. Genome Res, 2009, 19: 1953–1962
32 Berkofsky-Fessler W, Nguyen T Q, Delmar P, et al. Preclinical biomarkers for a cyclin-dependent kinase inhibitor translate to candidate pharmacodynamic biomarkers in phase I patients. Mol Cancer Ther, 2009, 8: 2517–2525
33 Ergun A, Lawrence C A, Kohanski M A, et al. A network biology approach to prostate cancer. Mol Syst Biol, 2007, 3: 82
34 di Bernardo D, Thompson M J, Gardner T S, et al. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. Nat Biotechnol, 2005, 23: 377–383
35 Ding L, Getz G, Wheeler D A, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature, 2008, 455: 1069–1075
36 Segal E, Friedman N, Koller D, et al. A module map showing conditional activity of expression modules in cancer. Nat Genet, 2004, 36: 1090–1098
37 Saijo T, Ishii G, Ochiai A, et al. Eg5 expression is closely correlated with the response of advanced non-small cell lung cancer to antimitotic agents combined with platinum chemotherapy. Lung Cancer, 2006, 54: 217–225
38 Okazaki Y, Furuno M, Kasukawa T, et al. Analysis of the mouse transcriptome based on functional annotation of 60770 full-length cDNAs. Nature, 2002, 420: 563–573
39 Metzker M L. Sequencing technologies—the next generation. Nat Rev Genet, 2010, 11: 31–46
40 Wang E T, Sandberg R, Luo S, et al. Alternative isoform regulation in human tissue transcriptomes. Nature, 2008, 456: 470–476
41 Mortazavi A, Williams B A, McCue K, et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods, 2008, 5: 621–628
42 Maher C A, Kumar-Sinha C, Cao X, et al. Transcriptome sequencing to detect gene fusions in cancer. Nature, 2009, 458: 97–101
43 Pan Q, Shai O, Lee L J, et al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet, 2008, 40: 1413–1415
44 Orom U A, Derrien T, Beringer M, et al. Long noncoding RNAs with enhancer-like function in human cells. Cell, 2010, 143: 46–58
45 Tripathi V, Ellis J D, Shen Z, et al. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol Cell, 2010, 39: 925–938

Supporting Information

File S1 BED file of lincRNAs. See Supplementary File 1.bed

File S2 List of functions of lincRNAs. See Supplementary File 2.xls

Table S1 Human paired-end RNA-Seq data sets from the Sequence Read Archive (SRA) database

Table S2 Gene expression datasets from GEO database used to construct 'two-color' co-expression network

Table S3 Properties of networks using different number of datasets

Table S4 Genetic mediators of protein coding genes

Table S5 LincRNA-sets dysregulated in lung cancer

Figure S1 Computational pipelines for identification of human lincRNAs.

Figure S2 Expression profiles of protein coding and lincRNA genes.

Figure S3 Genes differently expressed response to CDK inhibitor treatment.

Figure S4 Genomic contexts of three lincRNA genes.

Figure S5 The relationship between coding gene-centered gene sets and cancer.

Figure S6 An example of relationship between lincRNA-set and cancer.

The supporting information is available online at life.scichina.com and www.springerlink.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.